PARAMETERIZATION OF F0 REGISTER AND DISCONTINUITY TO PREDICT PROSODIC BOUNDARY STRENGTH IN HUNGARIAN SPONTANEOUS SPEECH

Uwe D. Reichel[1], *Katalin Mády*[2]

[1]*Institute of Phonetics and Speech Processing, University of Munich*
[2]*Research Institute for Linguistics, Hungarian Academy of Sciences*
reichelu@phonetik.uni-muenchen.de, mady@nytud.hu

Abstract: This study addresses the questions how to parameterize (1) aspects of fundamental frequency (F0) register, i.e. time-varying F0 level and range within prosodic phrases and (2) F0 discontinuities at prosodic boundaries in order to predict perceived prosodic boundary strength in Hungarian spontaneous speech. For F0 register stylization we propose a new fitting procedure for base-, mid-, and toplines that does not require error-prone local peak and valley detection and is robust against disturbing influences of high pitch accents and boundary tones. From these linear stylizations we extracted features which reflect F0 boundary discontinuities and fitted stepwise linear regression and regression tree models to predict perceived boundary strength. In a ten-fold cross-validation the mean correlation between predictions and human judgments amounts up to 0.8.

1 Introduction

1.1 Prosodic structure

Speech is prosodically structured into units that are separated by prosodic phrase boundaries. As shown in Figure 1A the time-varying intonation range within these prosodic phrases is given by a baseline and a topline that impose a lower and upper limit for local fundamental frequency (F0) movements [14]. These lines are defined by their F0 start points and their slopes. In declarative sentences baseline and topline usually have negative slopes and converge towards the end of the unit, which is referred to as declination [5, 8]. From this intonation range the intonation level can be inferred for example in form of a midline between base- and topline [9]. Both range and level are referred to as *register* in the literature [16].

The main phonetic correlates of the boundaries between these units are speech pauses [18], boundary tones [2], final lowering [10], pitch reset [6], prefinal lengthening [20], and a resistance against cross-boundary coarticulation [3]. [6] have demonstrated by perception experiments with delexicalized stimuli, that these acoustic features are also interpreted as boundary signals without any higher-level linguistic information.

Right-edge boundary tones mark whether or not the speaker intends to continue. Pitch reset serves to re-initialize the F0 register (level and range) to higher values after declination.

1.2 Parameterizations

Two questions are addressed: first, how to parameterize register within intonation units in a robust way, and second, how to parameterize the discontinuity signals at phrase boundaries.

Former studies typically address pitch range by fitting base- and toplines to local F0 minima and maxima by linear regression [11, 17]. The main shortcoming of this strategy consists in

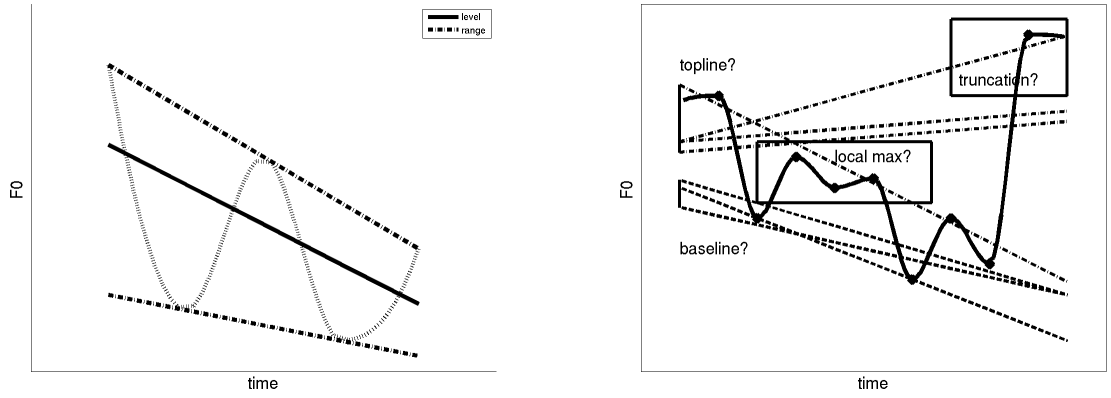

Figure 1 - **A (left):** F0 register as *level* and *range*. **B (right):** Problems of register stylization on the basis of local F0 peaks and valleys: fuzzy local peak detection and high dependency of the regression result on the choice of relevant peaks and valleys leads to 3 different baselines and 4 different toplines.

its vulnerability to large F0 displacements of boundary tones or prominent pitch accents, that highly disturb the F0 start point and the slope of the linear fit. One possible solution to remove the boundary tone disturbances consists in truncating the initial and final part of the contour, but as a consequence potentially relevant contour information will be ignored. Furthermore, we frequently experienced in the current study that after truncation the remaining contour is too short for a representative fit often resulting in a counter-intuitive crossing of the base- and the topline.

In Figure 1B some problems of automatic declination line fitting methods are summarized. Next to the fuzzy issue of local peak detection, the declination line offset and slopes highly depend on the choice of peaks to be relevant for the fit.

Alternatively, a more stable approach is given by fitting a midline to the whole F0 contour [11, 19]. This yields a more robust stylization but captures only the level but not the range aspect of declination. In section 3 we will introduce a fitting procedure, which is robust and which accounts for both intonation level and range.

In order to predict perceived prosodic boundary strength from the speech signal we propose several measures inferred from the F0 contour that reflect its discontinuity at boundaries.

2 Data

2.1 Corpus

The data comprises Hungarian spontaneous speech from maptask dialogues. In this study a corpus fragment consisting of 5 utterances of 10 speakers was analysed. It is manually segmented amongst others on the word level and contains prosodic boundary labels by 20 naive Hungarian subjects. The boundary label set comprises the tags *weak, strong* and *hesitation*. Hesitation and utterance-final boundary instances were discarded for the current analysis.

Boundary strength To avoid the strong assumption that boundary strength is perceived categorically, we transformed the categorical labeler judgments into a continuous measure of perceived strength ranging from 0 to 1. For this purpose we adopted the prominence score approach of [13], so that the perceived strength is given by the following formula: $\frac{2 \cdot n(s)+n(w)}{2 \cdot n(\text{subjects})}$, where

$n(s)$ and $n(w)$ stand for the number of *strong* and *weak* judgments respectively.

2.2 F0 Preprocessing

Within each segment voiceless parts are interpolated by piecewise cubic splines. F0 was then smoothed by Savitzky Golay filtering with a third order polynomial within a 5 sample window. For speaker normalization an F0 base value b was calculated as the median below the 5th percentile. F0 was then transformed to semitones relative to this base value as $F0_{st} = 12 \cdot log_2(\frac{F0_{Hz}}{b})$.

3 Register parameterization

At each word boundary the F0 segments of 1 second length preceding and following the boundary are taken for further analysis. The choice of 1 second is motivated by a trade-off that longer segments may contain more than one global declination event, and shorter segments may only contain local pitch events like pitch accents from which global declination cannot be inferred.
To capture F0 level and range we fitted a base-, a mid- and a topline to the F0 data. The fitting procedure consists of the following steps:

- A window of length 200 ms is shifted along the F0 vector with the stepsize of 10 ms.
- Within each window the F0 median is calculated
  - of the values below the 10th percentile for the baseline,
  - of the values above the 90th percentile for the topline, and
  - of all values for the midline.

  This gives 3 sequences of medians, one for the base-, the mid-, and the topline, respectively.
- Within each median sequence outliers are replaced by nearest neighbor interpolation.
- Finally, for all three median sequences linear polynomials are fitted whereas time is normalized to the interval $[0\ 1]$ so that the polynomial coefficients can directly be interpreted as F0 offset and normalized rate, respectively.

This procedure is shown in Figure 2A. Figure 2B shows the stylization results for the base-, mid-, and topline. It can be seen in this Figure, that using F0 medians relative to respective percentiles instead of local peaks and valleys makes the stylization robust against prominent pitch accents and boundary tones. Furthermore, it circumvents errors resulting from imperfect local peak detection.

4 Boundary parameterization

As pointed out in the introduction, prosodic boundaries are amongst others marked by boundary tones and discontinuity signals as pauses and pitch resets. Next to the pause length feature we derived several F0-related features capable to capture boundary tones and F0 discontinuity by the following parameterization:
Next to the separate register stylizations within 1 second intervals preceding and following a word boundary (s_1 and s_2, respectively) we carried out the same stylization for the concatenation

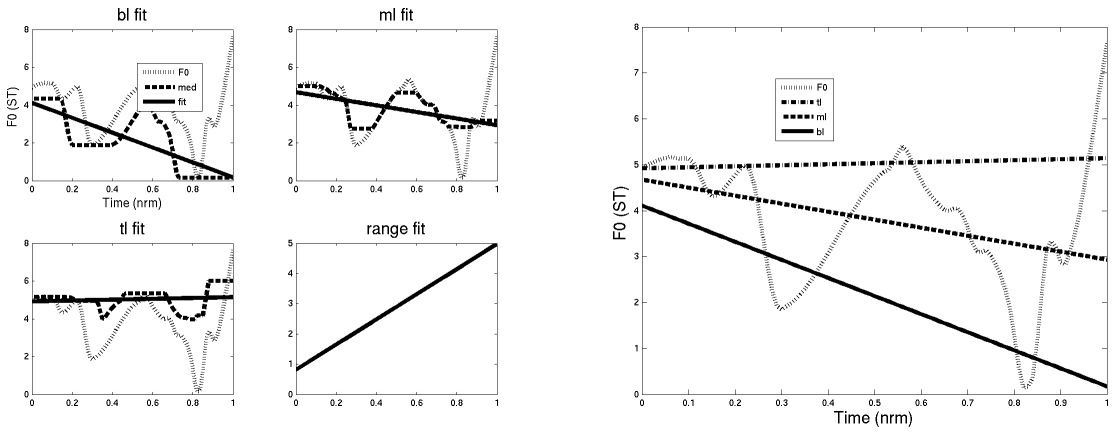

Figure 2 - **A (left):** Stylization of base-, mid- and topline based on F0 median sequences of different percentiles. Stylization of F0 range change. **B (right):** Line stylization result.

of these two intervals s_{12} which is shown for the baselines in Figure 3. The purpose of this threefold stylization is to infer boundary strength from declination line similarity. If s_1 and s_2 belong to the same prosodic phrase as is the case in Figure 3A, they are expected to have similar slopes and not to depart much from the s_{12} line in terms of offset and root mean squared deviation (RMS). Stronger prosodic boundaries in contrast are expected to be reflected by a lower degree of s_1-s_2 coupling. Such a discontinuity induced by pitch reset is expressed in clear offset and slope differences of the declination lines s_1 and s_2 and their larger deviations from s_{12} as can be seen in Figure 3B.

Boundary tones are modelled in 200 ms windows left- and right-adjacent to the word boundary in terms of the distances of the F0 medians to the corresponding top- and baseline F0 values. A median high above the topline indicates a high boundary tone, a median deep below the baseline a low tone. Both are considered as strong boundary markers.

In total, 60 features have been extracted for the segments s_1 preceding the word boundary, s_2 following the boundary, and their concatenation s_{12}. These features can be grouped as follows:

1. **pause length** was calculated in ms;
2. **level characteristics** were derived separately for base-, mid- and topline as the slope difference between s_1, s_2, and s_{12} and the RMS from s_1 and s_2 to s_{12}. Furthermore, the level reset from s_1 to s_2 in semitone (ST) has been extracted;
3. **range characteristics:** For s_1 and s_2 the global range characteristics were measured as the RMS between top- and baseline. Additionally, range changes were modelled in terms of linear regression slopes reflecting the time-varying distance between base- and topline. The range reset from s_1 to s_2 in ST was given by the difference of the final base-topline distance of s_1 and their initial distance in s_2;
4. **boundary tone:** the distances of the final median F0 to the final base- and topline F0 of s_1 were measured as well as the corresponding distances of the initial median F0 of s_2.

5 Prediction of perceived boundary strength

We trained and evaluated stepwise linear regressions and regression tree models [1] to predict perceived boundary strength. The incremental approach of the stepwise regression only incor-

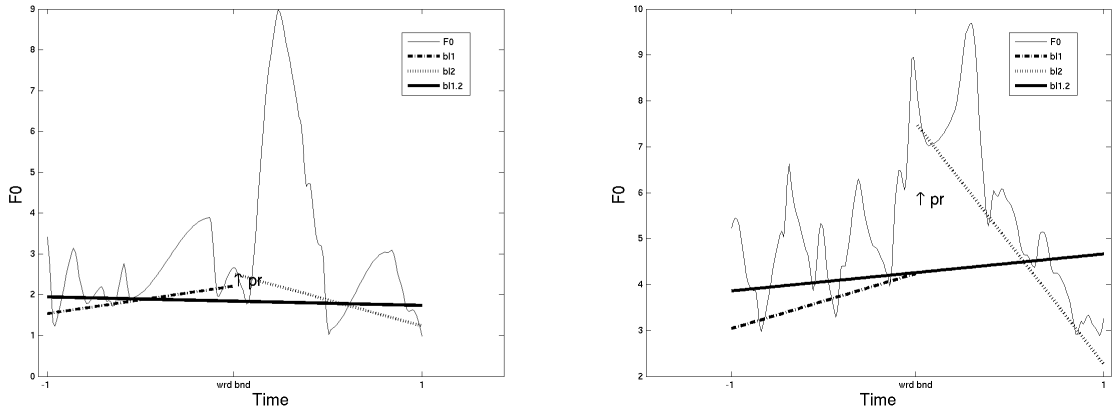

Figure 3 - **A (left):** No discontinuity indication (perceptual boundary judgment: 0). **B (right):** Strong boundary marked by high F0 discontinuity (judgment: 0.5). The parameterization adequately reflects that only the pitch reset (*pr*; B) but not the local pitch accent (A) influences the slope of *bl*2.

porating the features with the highest predictive power into the model accounts for the high number of extracted features in comparison to the available training data. For the regression trees we applied a sequential feature selection technique finding the optimal feature subset minimising the root mean squared error between prediction and human judgments in the training data.

In order to fulfill linearity, for the linear regression model, only absolute values were taken for each variable. All regression models were trained on z-scores. Furthermore, since the features were highly correlated with each other, we additionally trained the models on data orthogonalized by a principal component analysis.

6 Results

6.1 Feature impact

Table 1 shows a selection of the highest correlations of the perceived boundary strength to standard boundary features as well as new features derived from the discontinuity parameterization. It can be seen that the correlations of the new features are much lower than the correlations of the standard ones.

6.2 Boundary strength prediction

The correlations between human judgments and model predictions on all data and after a 10-fold cross-validation are shown in Table 2 and Figure 4, respectively. *LR* stands for stepwise linear regression and *TR* for tree regression. Reference models using only the pause length feature are marked by *-P*. Since feature orthogonalization did not improve performance, it is not plotted here. Correlations differed significantly (Kruskal-Wallis test χ^2_3, $p < 0.001$). Due to the small sample size a Tukey-Kramer post hoc comparison only revealed one significant difference between *LR* and *TR-P*. Nevertheless, it is quite apparent, that the models *LR* and *TR* using the intonation discontinuity features outperform the pause-only reference models with respect to the median correlation on the test data (> 0.8 as opposed to 0.3) and their robustness is expressed in a significantly lower performance variance on unseen data (F-test, $F[7,9] = 0.15$, $p < 0.05$).

Standard Features		New discontinuity features	
pause		**range**	
length	0.71	s_1 vs. s_{12} final RangeDiff	0.21
boundary tone		**level**	
s_1 blDist	0.54	s_1 vs. s_{12} tlOffsetDist	0.22
s_1 mlDist	0.54	s_1 vs. s_2 SlopeDiff	0.16
s_1 tlDist	0.40	s_{12} vs. $s_1.s_2$ tlRms	0.19

Table 1 - Correlations of standard and new discontinuity features to perceived boundary strength. s_1, s_2, s_{12}: F0 segments preceding and following the word boundary and their concatenation; *bl, ml, tl:* base-, mid- topline; **Dist:* distance of the boundary tone to the respective line in ST; *final RangeDiff*: range difference in ST between s_1 and s_{12} at the end of s_1; *OffsetDist* Distance in ST between the F0 starting points of the regression lines; *SlopeDiff:* difference of the declination slope coefficients; *Rms:* root mean squared deviation.

Model	r	RMSE
Linear Regression LR	0.89	0.08
Regression Tree TR	0.90	0.08
Baseline (pause only)	0.71	0.13

Table 2 - Model evaluation on all data.

7 Discussion

Register parameterization The parameterization approach proposed in this study does not require any detection of local peaks and valleys and is robust against local pitch events as is illustrated in Figure 2. It can be seen, that due to the high boundary tone, a simple fit through local F0 maxima would result in a steep topline collapsing with the baseline at the beginning of the segment, which lacks plausibility. An alternative truncation approach that cuts off the initial and final part of the contour runs the risk to remove relevant F0 information and often produces contours which are too short for a reliable declination estimation.

Figure 3 shows, that our stylization is capable to distinguish between prominent F0 movements of phrase-internal local events like the pitch accent in 3A from boundary signals like the pitch reset in 3B. This captured discontinuity difference is well reflected in the perceptual boundary strength scores, amounting 0 for 3A, and 0.5 for 3B.

Boundary parameterization Based on the register fittings within the segments before and after a word boundary, several discontinuity features have been proposed in order to quantify boundary strength. They are mainly based on examining the coupling of declination lines. The larger their difference concerning slope and offset, the more likely they belong to different prosodic phrases. Nevertheless, correlations between the new features and perceived boundary strength turned out to be rather low.

Predicting perceived boundary strength We accounted for the high feature number and the high inter-feature correlations by stepwise linear regression, sequential feature selection, and orthogonalization by principal component analysis. The linear regression model cannot cope with non-linear relationships between boundary variables and the target. Therefore features like a slope difference between two stylization lines can only be inserted as absolute values to the

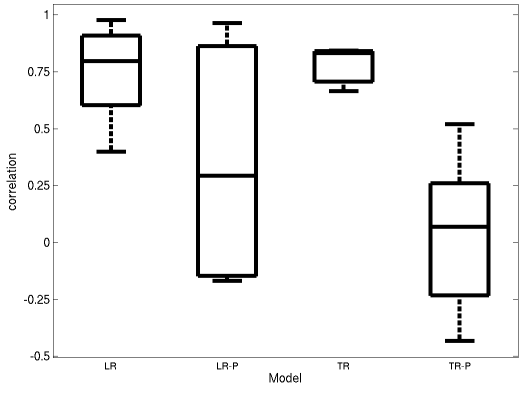

Figure 4 - Correlations between model predictions and perceived boundary strength after 10-fold cross-validation. LR-: stepwise linear regression, TR-: tree regression, -P: pause only.

model, since highly positive as well as negative values are expected to cause discontinuity. By doing this, potential differences related to the algebraic sign cannot be accounted for. Nevertheless, the linear regression models were not outperformed by the regression trees, which might indicate that the more complex tree models and the sequential feature selection procedure suffer from data sparseness. In any case, despite of the low correlations between the new features and boundary strength, they contribute in improving strength prediction models, amongst others because in Hungarian pauses are not exclusively connected to strong boundaries only [12], which weakens the reliability of this feature.

Applications The proposed register parameterization approach can be used for the decomposition of global and local intonation components in superpositional models as described in [7] and [15]. Hereby the local component's contribution could be expressed relative to the base- or midline, or could be normalized to the range between the base- and the topline.
Our boundary strength models can be adopted for automatic boundary labelling simply by transforming the continuous model output between 0 and 1 into a categorical decision in dependence on the application needs.

8 Acknowledgments

The work of the first author has been carried out within the CLARIN-D project [4] (BMBF-funded). The second author was funded by OTKA 101050 "A laboratory phonology approach to Hungarian prosody".

References

[1] BREIMAN, L., J. FRIEDMAN, C. STONE and R. OLSHEN: *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove, CA., 1984.

[2] BROWN, G., K. CURRIE and J. KENWORTHY: *Questions of Intonation*. Croom Helm, London, 1980.

[3] CHO, T.: *Prosodically-conditioned strengthening and vowel-to-vowel coarticulation*. Journal of Phonetics, 32:141–176, 2004.

[4] *http://eu.clarin-d.de/index.php/en/*. Clarin-D web page.

[5] COHEN, A., R. COLLIER and J. T'HART: *Declination: construct or intrinsic feature of speech pitch*. Phonetica, 39:254–273, 1982.

[6] DE PIJPER, J. and A. SANDERMANN: *On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues*. Journal of the Acoustical Society of America, 96:2037–2047, 1994.

[7] FUJISAKI, H.: *A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour*. In FUJIMURA, O. (ed.): *Vocal physiology: voice production, mechanisms, and functions*, pp. 165–175. Raven, New York, 1987.

[8] LADD, D.: *Declination: A review and some hypotheses*. Phonology Yearbook, 1:53–74, 1984.

[9] LADD, D.: *On the theoretical status of ïhe baselineïn modelling intonation*. Language and Speech, 36(4):435–451, 1993.

[10] LIBERMAN, M. and J. PIERREHUMBERT: *Intonational Invariance under Changes in Pitch Range and Length*. In ARONOFF, M. and R. OEHRLE (eds.): *Language Sound Structure*, pp. 157–233. MIT Press, Cambridge, MA, 1984.

[11] LIEBERMAN, P., W. KATZ, A. JONGMAN, R. ZIMMERMAN and M. MILLER: *Measures of the sentence intonation of read and spontaneous speech in American English*. Journal of the Acoustical Society of America, 77:649–657, 1985.

[12] MÁDY, K. and F. KLEBER: *Variation of pitch accent patterns in Hungarian*. In *Proc. Speech Prosody*, pp. 100924:1–4, Chicago, 2010.

[13] MO, Y., J. COLE and M. HASEGAWA-JOHNSON: *Prosodic effects on vowel production: evidence from formant structure*. In *Proc. Eurospeech*, pp. 2535–2538, Brighton, 2009.

[14] PIKE, K.: *The intonation of American English*, vol. 1 of *University of Michigan publications*. University of Michigan Press, Ann Arbor, 1945.

[15] REICHEL, U.: *The CoPaSul intonation model*. In KROEGER, B. and P. BIRKHOLZ (eds.): *Elektronische Sprachverarbeitung 2011*, Studientexte zur Sprachkommunikation, pp. 341–348. TUDpress, 2011.

[16] RIETVELD, T. and P. VERMILLION: *Cues for Perceived Pitch Register*. Phonetica, 60:261–272, 2003.

[17] SCHMID, C. and M. GENDROT, C. ADDA-DECKER: *Une comparaison de déclinaison F0 entre le français et l'allemand journalistiques*. In *JEP-TALN-RECITAL*, vol. 1, pp. 329–336, 2012.

[18] SWERTS, M. and R. GELUYKENS: *Prosody as a marker of information flow in spoken discourse*. Language and Speech, 37(1):21–43, 1994.

[19] SWERTS, M., E. STRANGERT and M. HELDNER: *F0 declination in read-aloud and spontaneous speech*. In *Proc. ICSLP*, vol. 3, pp. 1501–1504, Philadelphia, 1996.

[20] WIGHTMAN, C., S. SHATTUCK-HUFNAGEL, M. OSTENDORF and P. PRICE: *Segmental Durations in the Vicinity of Prosodic Phrase Boundaries*. JASA, 91(3):1707–1717, 1992.